Behavior-Aware Network Segmentation using IP Flows

The 14th International Conference on Availability, Reliability and Security

August 26 – August 29, 2019

University of Kent, Canterbury, UK

Juraj Smeriga, **Tomas Jirsik**

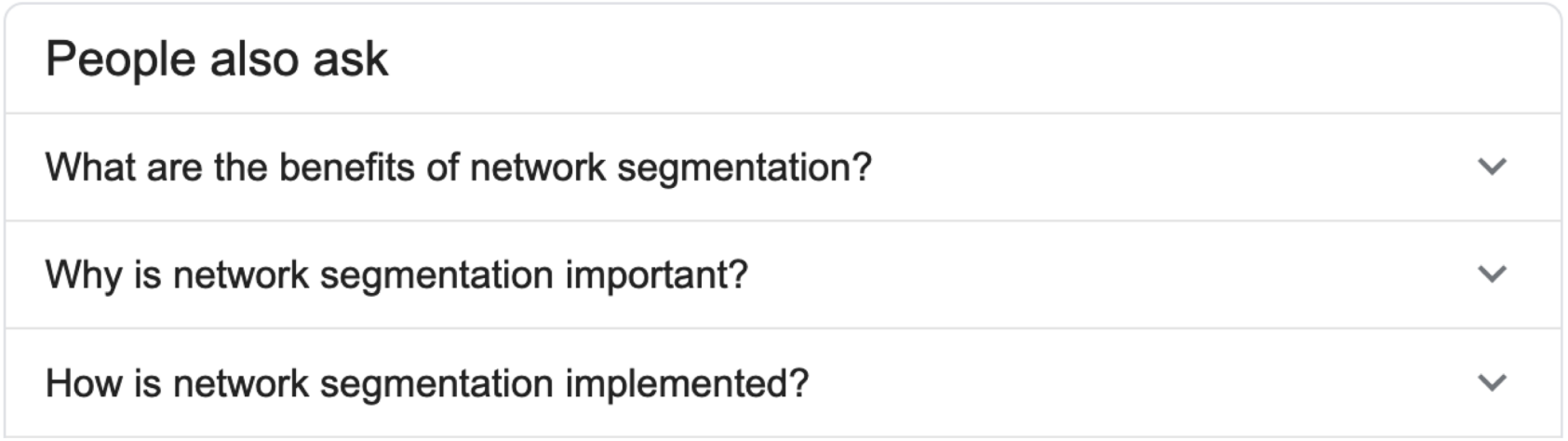
Institute of Computer Science,
Masaryk University, Czech Republic

CSIRT-MU

Network Segmentation

What is it good for?

Network segmentation *in computer networking is the act or* ***practice of splitting*** *a computer network into* ***subnetworks****, each being a* ***network segment****.*



Network IP Flow Monitoring

IP flows tell the stories

Connection-oriented network traffic observation

- Aggregates packets by flow keys
- Optimized for high speed, large-scale networks
- **Who** is communicating with **whom**, how **long**, on **which port/protocol**
- Application protocols monitoring – HTTP, DNS

Flow start	Duration	Proto	Src IP Addr:Port	->	Dst IP Addr:Port	Flags	Packets	Bytes
09:41:21.763	0.101	TCP	172.16.96.48:15094	->	209.85.135.147:80	.AP.SF	4	715
09:41:21.893	0.031	TCP	209.85.135.147:80	->	172.16.96.48:15094	.AP.SF	4	1594

What is the Problem?

Problems, problems everywhere

- **Complexity of networks** – multilayered network, dynamics
- **Lack of information** – limited/no access to all hosts in a network
- **Connection-oriented IP Flows** – host-oriented view is required
- **Large volume of data** – impossible to process manually

What are the segments?

How to assign hosts to the segments?

What is the Problem?

Problems, problems everywhere

Machine learning solves it all

What is the Problem?

Problems, problems everywhere

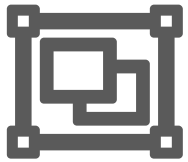
Machine learning solves it all

Really?

Hypotheses

Choosing the right question.

*Explore the **possibilities** of utilizing **machine learning** on **IP flows** to create **behavior-consistent** network segments.*

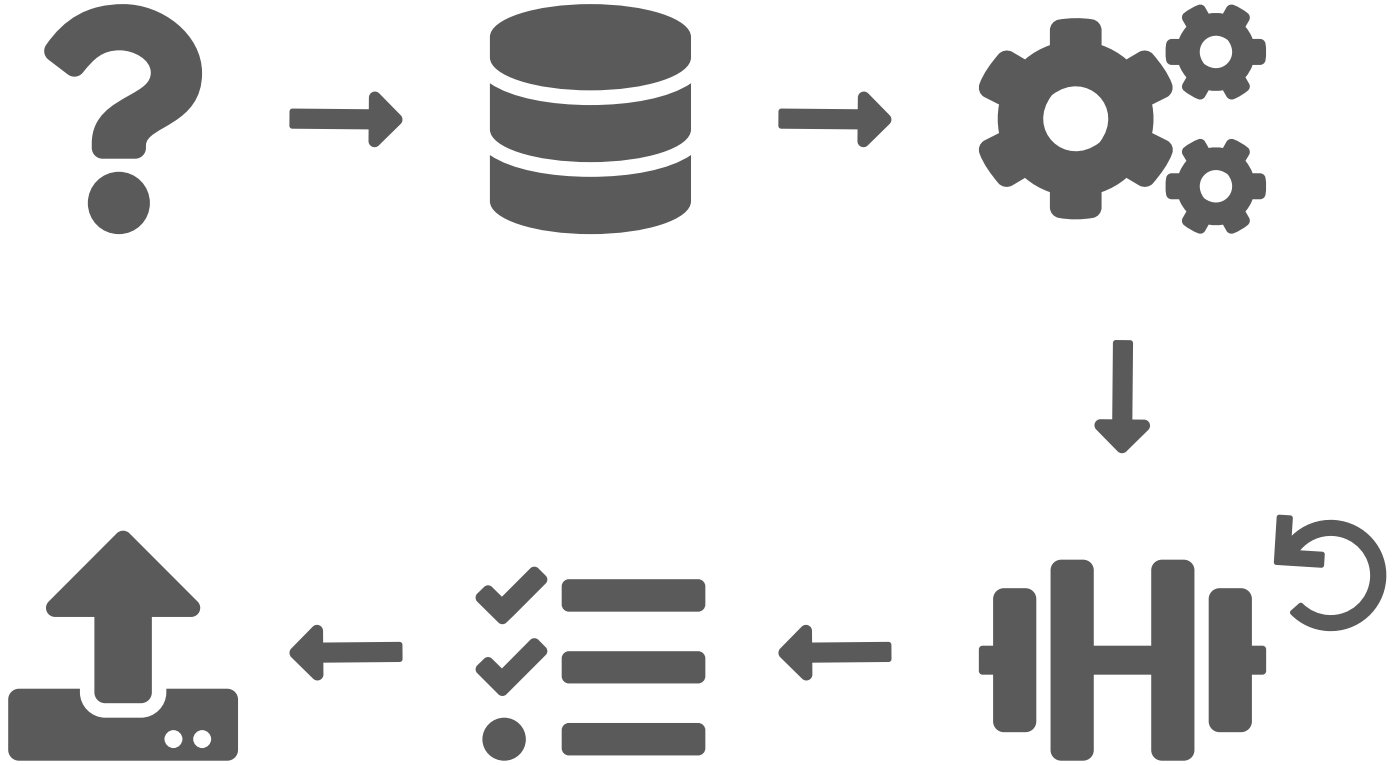


Network can be divided into behavior-consistent segments using machine learning.

It is possible to assign an unknown host to an existing segment based on its behavior.

Methodology

It’s about the journey, not the destination



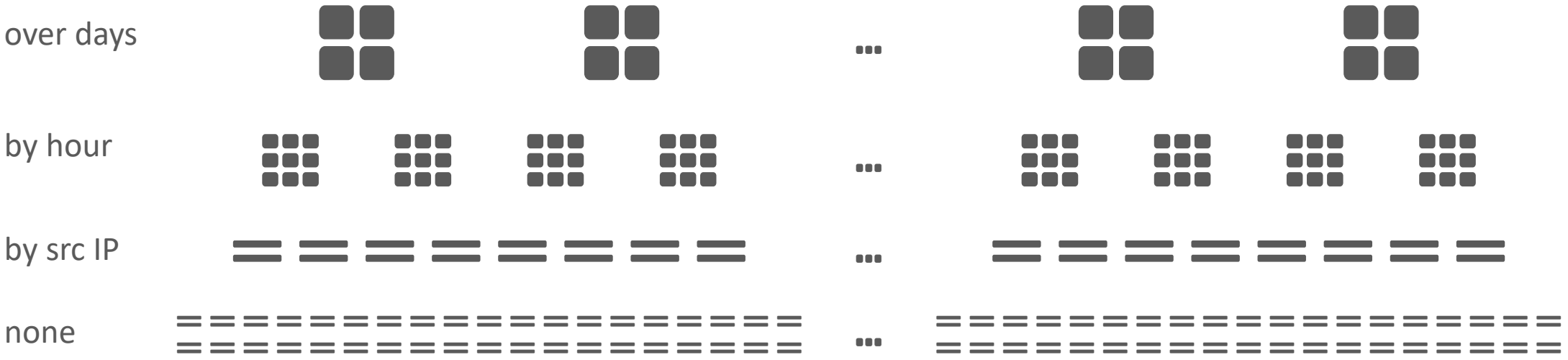
Dataset

Data collection

From connections to host profiles

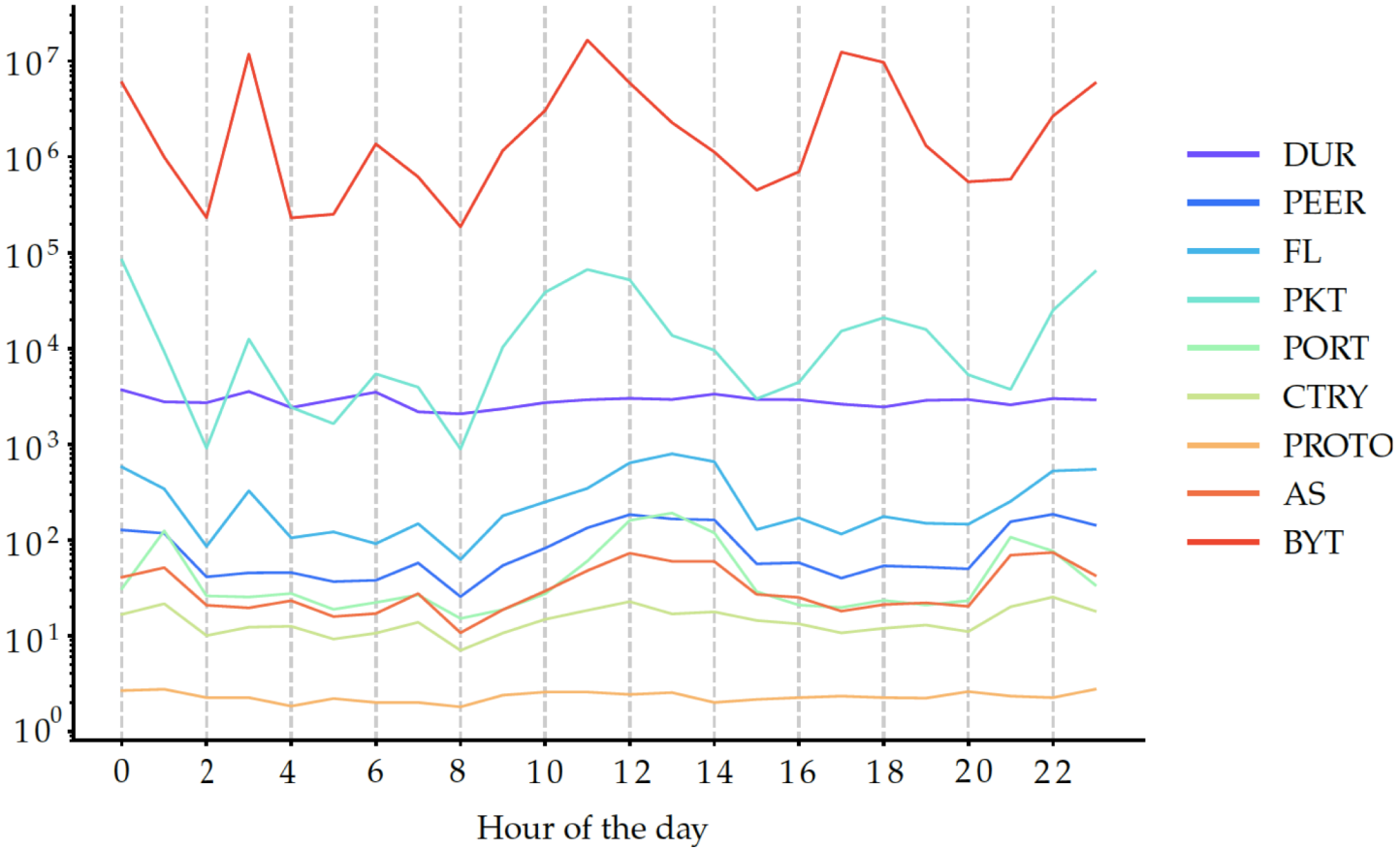
Features

- **1 month of data from /16 campus network**
- **Aggregations** – flow duration, number of packets, bytes, flows
- **Distinct counts** – peers, ports, protocols, AS numbers, country



Data collection

From connections to host profiles



Dataset

No more ”garbage in, garbage out”

Labelling

- **Origin** – list of existing administrative units (network ranges)
- **Labels** – range, administrative unit, and administrative subunit

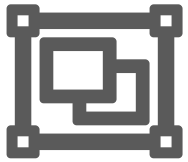
Preprocessing

- **Missing Values** – missing labels (9.18%), all missing values (42.74%), other replaced by 0, remains 31 501 hosts
- **Outliers** – 0.95 quantile
- **Standardization** – zero mean and unit variance
- **Dataset balancing** – undersampling of the major unit by 75%

Release

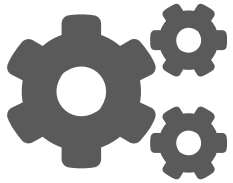
- **Anonymization** – IP addresses and ranges anonymized by CryptoPan
- **Publishing platform** – zenodo.org with feature description

Network Segment Discovery



Algorithms

Clustering – identifying groups in unknown

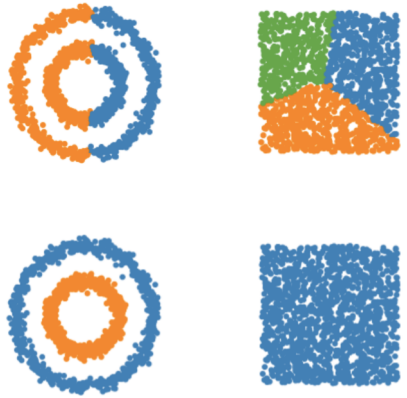


What class of algorithm?

- **Problem** – divide hosts into a previously unknown groups of similar hosts
- **Unsupervised ML - Clustering Algorithms** - the task of grouping a set of objects in such a way that objects in the same are more similar to each other than to those in other groups

Selected Clustering Algorithms

- **K-Means**
  - 👍 simple, fast, scales to large datasets
  - 👎 predefined number of clusters, initial centroids matters, curse of dimensionality
- **Density-based spatial clustering of applications with noise (DBSCAN)**
  - 👍 no need for predefined number of clusters, non-convex cluster identification
  - 👎 non-determinism, heavy dependence on selected distance measure
- **Time-series modification**
  - **LB Keogh Dynamic time warping** instead Euclidean distance



Training

Practice makes perfect

K-Means

- **Number of clusters** – 22 equal to number of administrative units
- **Initial centroids** – random selection
- **Max iterations** – 300

DBSCAN

- **Elbow identification** – minPts = 44, ε = 160
- **Grid search** – minPts = 40, ε = 5

Evaluation

- **Silhouette coefficient** – no labels, <-1 *(bad)*, 1 *(good)*>,
- **Adjusted Rand index** – labels, around 0 *(bad)*, 1 *(good)*

Results

Are there behavior-consistent segments?

Number of clusters

- DBSCAN optimum 7 clusters

Initial Results

Algorithm	Silhouette	ARI
K-Means	0.02	0.30
K-Means_LB_K-DTW	0.07	0.16
DBSCAN	0.08	-0.13

Takeaway

- **A less behavior-similar segments** than the administrative ones
- Segments are **overlapping**
- DBSCAN is **slightly better** for clustering behaviors on network

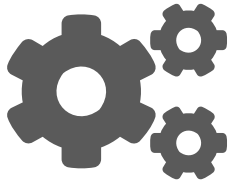
Advanced Analysis

Segment	Algorithm	Silhouette	ARI
Reduced	K-Means	0.001	0.92
	K-Means_LB_K-DTW	0.13	0.22
	DBSCAN	0.19	-0.04
Binary	K-Means	0.001	0.95
	K-Means_LB_K-DTW	0.03	0.13
	DBSCAN	0.001	-0.14

Network Segment Assignment

Algorithms

Classification – assigning to a category



What class of algorithm?

- **Problem** – assign a new host into an existing segment
- **Supervised ML - Classification Algorithms** – based on the data creates model and predict the class of given data points

Selected Classification Algorithms

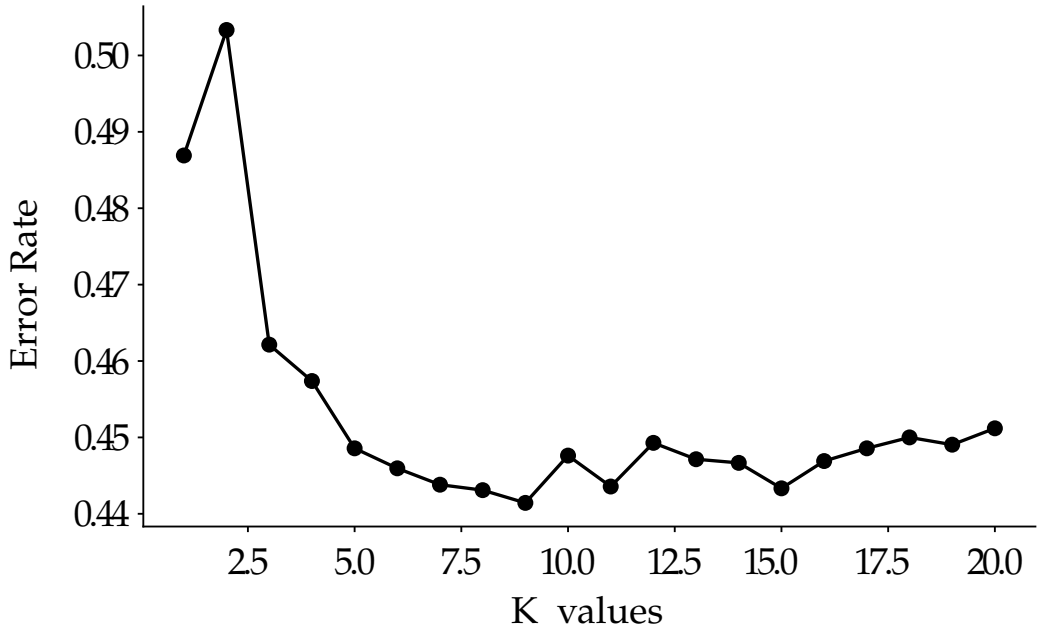
- **K-nearest neighbors**
  - 👍 simple, only one parameter
  - 👎 homogenous features, curse of dimensionality
- **Support Vector Machines**
  - 👍 kernel choice, avoids overfitting
  - 👎 plenty of parameters to set
- **Decision Trees**
  - 👍 easy to understand, requires little data preparation
  - 👎 non-robust, overfitting

Training

Practice makes perfect

K-nearest neighbors

- ***k* value setting** – elbow analysis



SVM

- **Kernel** – polynomial
- **Penalty parameter, kernel coef.** – grid search
  - Penalty parameter – 0.01
  - Kernel coef. – 1
- **Uniform weights, no iteration limit**

Decision Trees

- **Split –** Gini impurity
- **Max features considered –** 22
- **No depth limit**

Evaluation

- **Train : test ratio** – 80:20, random selection
- **Metrics** – precision, recall, F-Score

Results

Is it possible to assign a host?

Initial Results

Algorithm	Precision	Recall	F-Score
KNN	0.56	0.52	0.52
KNN_LB_K-DTW	0.40	0.41	0.36
SVM	0.62	0.62	0.60
DTs	0.61	0.61	0.61

Advanced Analysis

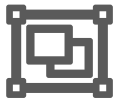
Segment	Algorithm	Precision	Recall	F-Score
Reduced	KNN	0.75	0.74	0.74
	SVM	0.82	0.81	0.81
	DTs	0.78	0.78	0.78
Binary	KNN	0.87	0.87	0.87
	SVM	0.90	0.90	0.90
	DTs	0.92	0.92	0.92

Takeaway

- Noise is introduced by **small fuzzy administrative** segments
- Hosts with similar behaviors are **present in more** administrative **segments**
- **DT** and **SVM** performs better than KNN
- **No time causality** required for classification

Conclusions

Take away messages



We can divide network to behavior-consistent segments using ML

- **A less behavior-similar segments** than the administrative ones
- Segments are **overlapping**
- DBSCAN is **slightly better** for clustering behaviors on network

It is possible to assign an unknown host to an existing segment based on its behavior.

- Noise is introduced by **small fuzzy administrative** segments
- Hosts with similar behaviors are **present in more** administrative **segments**
- **No time causalit**y required for classification
- **DT** and **SVM** performs better than KNN

Summary

Our contributions

Creation of dataset with features suitable for host behavior modelling

Identification what ML techniques can be used for behavior-aware network segmentation

Comparison of the performance of the ML techniques

Experiment and data released for public use

Experiment Download:

https://github.com/CSIRT-MU/BehaviorNetworkSegmentation

https://csirt.muni.cz

@csirtmu

Tomas Jirsik

jirsik@ics.muni.cz

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

MUNI
ICS

CSIRT-MU